Modèle de langue visuel pour la reconnaissance de scènes

Trong-Ton Pham, Loic Maisonnasse, Philippe Mulhem, Eric Gaussier

Modèle de langue visuel pour la reconnaissance de scènes

Trong-Ton Pham[1,2]**, Loïc Maisonnasse**[3]**, Philippe Mulhem**[1]**, Eric Gaussier**[1]

[1] *Laboratoire d'Informatique de Grenoble UJF-CNRS - 38041 Grenoble Cedex 9, France philippe.mulhem@imag.fr, eric.gaussier@imag.fr*
[2] *IPAL-I2R, Fusionopolis, Singapore 138632 - ttpham@i2r.a-star.edu.sg* [3] *Université de Lyon, INSA-Lyon, LIRIS loic.maisonnasse@imag.fr*

RÉSUMÉ. *Dans cet article, nous décrivons une méthode pour utiliser un modèle de langue sur des graphes pour la recherche et la catégorisation d'images. Nous utilisons des régions d'images (associées automatiquement à des concepts visuels), ainsi que des relations spatiales entre ces régions, lors de la construction de la représentation sous forme de graphe des images. Notre méthode gère différents scénarios, selon que des images isolées ou groupées soient utilisés comme base d'apprentissage ou de tests. Les résultats obtenus sur un problème de catégorisation d'images montre (a) que la procédure automatique qui associe les concepts à une image est efficace, et (b) que l'utilisation des relations spatiales, en plus des concepts, permet d'améliorer la qualité de la classification. Cette approche présente donc une extension du modèle de langue classique en recherche d'information pour traiter le problème de recherche et de catégorisation d'images représentées par des graphes sans se préoccuper des annotations d'images.*

ABSTRACT. *We describe here a method to use a graph language modeling approach for image retrieval and image categorization. Since photographic images are 2D data, we first use image regions (mapped to automatically induced concepts) and then spatial relationships between these regions to build a complete image graph representation. Our method deals with different scenarios, where isolated images or groups of images are used for training or testing. The results obtained on an image categorization problem show (a) that the procedure to automatically induce concepts from an image is effective, and (b) that the use of spatial relationships, in addition to concepts, for representing an image content helps improve the classifier accuracy. This approach extends the language modeling approach to information retrieval to the problem of graph-based image retrieval and categorization, without considering image annotations.*

MOTS-CLÉS : *Représentation de graphes, recherche d'image, catégorisation d'image*

KEYWORDS: *Graph representation, image retrieval, image categorization*

1. Introduction

Après presque 20 ans de recherche dans le domaine de la recherche d'images, ce sujet est toujours considéré comme un défi pour les chercheurs. Les problèmes auxquels ce domaine est confronté ont trait au fossé sémantique ainsi qu'à la manière de représenter le contenu des images. Un autre élément important est qu'il n'est pas rare qu'une image soit reliée à d'autres images. Par exemple, tous les appareils de photographie numériques incorporent l'heure et la date de prise de vue, et cette information peut être utilisée pour les grouper (Platt *et al.,* 2003). De plus, les informations de géolocalisation, qui tendent à se généraliser, peuvent également être utiles (Kennedy *et al.,* 2007). Les groupements d'image peuvent alors profiter à l'indexation et la recherche d'images. Dans ce cas, il faut intégrer des moyens de faire correspondre des groupes. Nous montrons que l'utilisation de modèles de langue peut aisément s'appliquer à des groupes d'images requêtes et documents, et qu'une telle approche est robuste par rapport aux différences entre ces groupements.

Plusieurs travaux ont par le passé proposé l'utilisation de relations spatiales entre régions d'image pour leur indexation et leur recherche. Par exemple, les descriptions par chaînes 2D (2D strings) comme on les trouve dans le système Visualseek (Smith *et al.,* 1996) capturent les séquences d'apparition d'objets suivant une ou plusieurs directions de lecture. Cependant, la recherche de telles chaînes est complexe car elle est basée sur des recherches de sous-chaînes, ce qui est coûteux. Même si des heuristiques ont été proposées, comme dans (Chang *et al.,* 2000), afin d'accélérer (d'un facteur 10) le temps de calcul. D'autres travaux ont considéré l'utilisation de régions d'images dans des modèles probabilistes, en se basant par exemple sur des modèles de Markov cachés 1D (Iyengar *et al.,* 2005) ou 2D, comme dans (Smith *et al.,* 1996) et (Yuan *et al.,* 2007). Ces travaux s'intéressent à l'annotation d'images et n'utilisent pas les relations lors du traitement des requêtes. Les relations entre des éléments d'images peuvent également être exprimés par l'intermédiaire de conventions de nommage, comme dans (Papadopoulos *et al.,* 2007) où les relations sont utilisées pour l'indexation. Enfin des travaux tels que (Mulhem *et al.,* 2006) se sont focalisés sur des graphes conceptuels pour l'indexation et la recherche des images. Les représentations explicites de relations provoquent la génération de graphes complexes, ayant un impact négatif sur la correspondance de graphe qui est déja coûteuse (Ounis *et al.,* 1998). Un aspect de notre travail est de représenter le contenu des images par des graphes, sans souffrir du poids de la complexité de la correspondance de graphes durant l'étape de recherche. Pour cela, nous proposons de nous appuyer sur un ensemble de travaux existants dans le domaine de la recherche d'information.

L'approche à base de modèles de langue pour la recherche d'information existe depuis la fin des années 90 (Ponte *et al.,* 1998). Dans ce cadre, la valeur de pertinence d'un document pour une requête donnée est estimée par la probabilité que la requête soit générée par le document. Même si cette approche a été initialement proposée pour des unigrammes (c'est-à-dire des termes isolés), plusieurs extensions ont été proposées pour traiter des *n-grammes* (i.e. des déquences de termes) (Song *et al.,* 1999, Srikanth *et al.,* 2002), et plus récemment, des relations entre termes et égale-

ment des graphes. Par exemple, (Gao *et al.,* 2004) propose a) d'utiliser un analyseur de dépendance pour représenter les documents et les requêtes, et b) une extension de l'approche à base de modèle de langue pour manipuler ces arbres. Maisonnasse *et al.* (Maisonnasse *et al.,* 2007, Maisonnasse *et al.,* 2008) ont étendu cette approche avec un modèle compatible avec des graphes plus généraux, comme ceux obtenus par une analyse conceptuelle des documents et des requêtes. D'autres approches (comme (Fergus *et al.,* 2005, Gosselin *et al.,* 2007)) ont respectivement utilisé des réseaux probabilistes et des noyaux pour capturer des relations dans les images, ce qui est également notre intention ici. Dans le cas de (Fergus *et al.,* 2005), l'estimation des probabilités des régions repose sur l'algorithme EM, qui est sensible aux probabilités initiales. Dans le modèle que nous proposons, au contraire, la fonction de vraisemblance est convexe et possède un maximum global. Dans le cas de (Gosselin *et al.,* 2007), le noyau utilisé ne considère que les trois plus proches régions d'une région de référence. Nous intégrons dans notre modèle toutes les régions voisines d'une région. Enfin, contrairement à ces travaux, à ceux de (Iyengar *et al.,* 2005) et de (Barnard *et al.,* 2003) basés sur des modèles de langues, nous utilisons explictement des étiquettes de relations spatiales. Nous étendons en fait ici le travail de (Maisonnasse *et al.,* 2007) en appliquant, d'une part, ce travail à des images, et en considérant, d'autre part, que les concepts et les relations peuvent être pondérés.

La suite de cet article est organisé comme suit : la section 2 présente le modèle de langue visuel (VLM) utilisé pour décrire le contenu des images, ainsi que la procédure de correspondance utilisée pour calculer la similarité entre images ; la section 3 décrit ensuite les résultats obtenus par notre approche pour un problème de catégorisation portant sur 101 classes ; nous concluons en section 4.

2. Le modèle de langue pour les graphes d'images

2.1. *Modélisation des images avec des graphes visuels*

Notre objectif ici est de générer automatiquement, à partir d'une image donnée, un graphe qui représente son contenu. Un tel graphe contient les concepts associés à des éléments présents dans l'image, ainsi que les relations qui dénotent comment les concepts sont reliés dans l'image. Pour cela, notre procédure est basée sur quatre étapes :

1) Identifier les régions de l'image qui vont former les blocs de base pour l'identification de concepts.

2) Indexer chaque région avec un ensemble prédéfini de caractéristiques.

3) Regrouper toutes les régions de la collection en K classes, chaque classe représentant un concept. A la fin de cette étape, chaque région de l'image est représentée par un concept, qui est le nom de la classe à laquelle la région appartient. L'ensemble des concepts, que nous notons $\mathcal{C}$, correspond donc à l'ensemble des classes obtenues.

4) Extraire enfin les relations entre les concepts.

La première étape, l'identification de régions, peut être basée sur un découpage arbitraire de blocs non recouvrants de taille égale (par exemple en divisant une image en 25 blocs, soit une division 5x5), ou bien sur des régions définies à partir de points d'intérêts (comme avec des points SIFT)[1]. Le second point vise à représenter les régions par des vecteurs afin de les regrouper. Les caractéristiques que nous avons retenues dans cet article sont des couleurs dans l'espace HSV, qui peuvent être extraites facilement et rapidement. Notre approche se base sur les K-moyennes pour la troisième étape, approche standard, mais d'autres méthodes sont également possibles. Enfin, la quatrième étape génère un ensemble de concepts associés par des relations. Nous nous concentrons ici sur les relations spatiales *au_dessus* et *à_gauche*. A la fin du processus complet, nous obtenons un ensemble de concepts reliés pour représenter une image. Il est à noter qu'un même concept peut apparaître plusieurs fois dans une image (quand différentes régions sont assignées à une même classe, comme cela arrive souvent pour des régions décrivant le ciel par exemple). Chaque concept est donc associé à une pondération qui dénote son nombre d'occurrences dans l'image. De même, chaque relation est associée à un poids dénotant le nombre de fois où elle est observée entre deux concepts donnés d'une image.

Dans la suite, nous désignerons l'ensemble des concepts pondérés qui décrivent une image par W_C. W_C est défini sur $\mathcal{C} \times \mathbb{N}$. Chaque association entre deux concepts c et c' est orientée (comme le sont les relations spatiales dans les travaux autour d'Acemedia (Papadopoulos *et al.*, 2007) par exemple), et est représentée par un triplet de la forme $< (c, c'), l, n(c, c', l) >$, où l est une étiquette de l'ensemble $\mathcal{L}$ des étiquettes possibles et $n(c, c', l)$ un entier. Un tel triplet s'interprète comme le fait qu'il existe dans l'image $n(c, c', l)$ relations portant l'étiquette l entre les deux concepts c et c'. Cette représentation permet de rendre compte de l'absence de relations entre deux concepts par la prise en compte d'une étiquette particulière. Les étiquettes que nous considérons dans la suite sont *au_dessus* et *à_gauche*, les relations inverses (*au_dessous* et *a_droite*) étant implicitement prises en compte dans la mesure où les relations sont orientées.

En résumé, un graphe représentant une image i est défini par $G =< W_C^i, W_E^i >$, avec :

$$\begin{aligned} W_C^i &= \{(c, n(c; i)), c \in \mathcal{C}\} \\ W_E^i &= \{((c, c'), l, n(c, c', l; i)), (c, c') \in \mathcal{C}^2, l \in \mathcal{L}\} \end{aligned}$$

$n(c; i)$ et $n(c, c', l; i)$ sont les poids et correspondent ici à des nombres d'occurrences.

Nous décrivons maintenant sur un exemple un tel graphe par la figure 1. Dans cette figure, considérons une image photographique I découpée en 9 blocs rectangulaires de tailles égales. Chaque bloc est associé à une étiquette élément de $\mathcal{C}$. On remarque que le concept c1 apparaît 3 fois dans l'image, ce qui amène dans la description du contenu par le graphe en bas de la figure à la notation "c1, 3". La même approche est utilisée pour obtenir $W_C^I = \{(c1, 3), (c2, 3), (c3, 2), (c4, 1)\}$. Nous utilisons les relations *à*

1. Dans ce dernier cas, les regions peuvent se recouvrir.

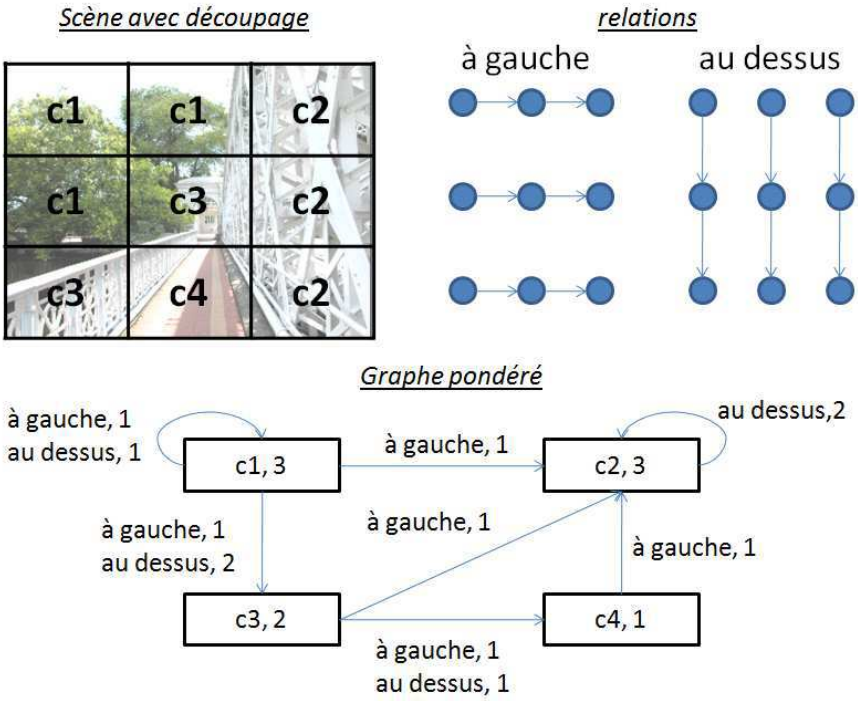

Figure 1. *Exemple de relations spatiales extraites d'une scène, avec le graphe correspondant.*

gauche et *au dessus* entre les régions, comme décrit dans la figure, afin de déterminer par exemple qu'entre les concepts c3 et c4 il existe une occurrence de la relation *à gauche*, donc $(c3, c4, \grave{a}\, gauche, 1) \in W_E^I$ et une occurrence de la relation *au dessus*, donc $(c3, c4, au\, dessus, 1) \in W_E^I$. Ce décompte se retrouve dans la description du graphe.

2.2. *Un modèle de langue pour les graphes visuels*

Notre fonction d'appariement entre images est fondée sur l'approche modèle de langue ((Ponte *et al.*, 1998)), étendue de façon à prendre en compte les éléments définis ci-dessus. De façon à différencier les images que l'on apparie, nous désignerons l'une de ces images par *image requête* et l'autre par *image document*.

La probabilité que le graphe d'une image requête $G_q (=< W_C^q, W_E^q >)$ soit généré à partir du graphe de l'image document G_d est définie par :

$$P(G_q|G_d) \quad = \quad P(W_C^q|G_d) \times P(W_E^q|W_C^q, G_d) \qquad [1]$$

Pour le premier terme du membre droit de l'équation ci-dessus, qui correspond à la probabilité de générer les concepts de l'image requête à partir du graphe de l'image document, nous nous reposons sur une hypothèse d'indépendance conditionnelle entre concepts, hypothèse classique en recherche d'information et en classification. La prise en compte des poids des concepts (c'est-à-dire, ici, des nombres d'occurrences des concepts) conduit naturellement à un modèle multinomial :

$$P(W_C^q|G_d) \quad \propto \quad \prod_{c \in \mathcal{C}} P(c|M_d)^{n(c;q)}$$

où $n(c;q)$ représente le nombre de fois que le concept c apparaît dans le graphe de l'image requête. Les paramètres du modèle $P(c|G_d)$ sont estimés par maximum de vraisemblance, avec un lissage de Jelinek-Mercer :

$$P(c|M_d) \quad = \quad (1-\lambda_u)\frac{F_1^d(c)}{F_1^d(.)} + \lambda_u \frac{F_1^{\mathcal{D}}(c)}{F_1^{\mathcal{D}}(.)}$$

où $F_1^d(c)$ représente le poids de c dans le graphe de l'image document et $F_1^d(.) = \sum_c F_1^d(c)$. Les fonctions $F_1^{\mathcal{D}}$ sont similaires, mais définies sur la collection, c'est-à-dire sur l'union des graphes des images de la collection. Le paramètre λ_u ($0 < \lambda_u < 1$) est le paramètre de lissage. Il joue le rôle d'un IDF (*Inverse Document Frequency*) ((Zhai *et al.*, 2004)) et permet de corriger une information peu fiable au niveau de l'image document par une information plus sûre extraite de la collection. Ce paramètre est en général réglé expérimentalement sur un ensemble d'apprentissage.

En suivant un processus similaire pour les relations, nous obtenons :

$$P(W_E^q|W_C^q, G_d) \quad \propto \quad \prod_{(c,c',l) \in \mathcal{C}^2 \times \mathcal{L}} P(L(c,c') = l|W_C^q, G_d)^{n(c,c',l,q)} \qquad [2]$$

où $L(c,c')$ est une variable à valeurs dans $\mathcal{L}$ qui rend compte des étiquettes possibles entre c et c'. Comme précédemment, les paramètres du modèle $P(L(c,c') = l|W_C^q, G_d)$ sont estimés par maximum de vraisemblance avec un lissage de Jelinek-Mercer, ce qui donne :

$$P(L(c,c') = l|W_C^q, G_d) = (1-\lambda_e)\frac{F_2^d(c,c',l)}{F_2^d(c,c',.)} + \lambda_e \frac{F_2^{\mathcal{D}}(c,c',l)}{F_2^{\mathcal{D}}(c,c',.)}$$

où $F_2^d(c,c',l)$ représente le nombre de fois que les concepts c et c' sont reliés par l'étiquette l dans le graphe de l'image document et $F_2^d(c,c',.) = \sum_{l \in \mathcal{L}} F_2^d(c,c',l)$. Par convention, dans le cas où l'un des deux concepts n'apparaît pas dans le graphe de l'image document :

$$\frac{F_2^d(c,c',l)}{F_2^d(c,c',.)} = 0$$

Les fonctions $F_2^{\mathcal{D}}$ sont similaires mais définies sur toute la collection (i.e., comme précédemment, sur l'union des graphes des images de la collection).

Le modèle que nous venons de présenter est inspiré du modèle défini dans (Maisonnasse *et al.,* 2008). Il en diffère cependant car (a) nous proposons ici une méthodologie complète de représentation d'une image à un niveau que nous qualifions de *conceptuel*, et (b) nous considérons ici des poids sur chaque concept et chaque relation. Dans la mesure où les poids que nous avons considérés sont des entiers, nous nous sommes reposés sur des distributions multinomiales pour modéliser l'appariement entre graphes. Des poids réels conduiraient à la considération d'autres types de distributions (par exemple de type Dirichlet). Nous allons maintenant illustrer le comportement de notre modèle dans le cadre de la classification d'image.

3. Expérimentations

Nous montrons ici la validité de notre approche dans le cadre d'une tâche de classification d'images. Plus précisément, nous voulons vérifier que a) notre proposition d'indexation conceptuelle est bien fondée, et b) que les relations spatiales ont un impact positif dans la caractérisation du contenu des images que nous proposons. Nous mettons également en évidence que notre méthodologie est robuste par rapport aux changements de scénarios présentés.

3.1. *La collection STOIC-101*

La collection *Singapore Tourist Object Identification Collection* est une collection d'images de 101 lieux d'intérêt touristique de Singapour (majoritairement des photographies d'extérieur). Ces localisations peuvent être vues comme des classes pour chacune des 3849 images. Ces images ont été prises dans leur majorité par des appareils photographiqes numériques, de manière similaire à des touristes, de 3 distances et 4 angles différents, avec des occlusions ou des cadrages partiels, de façon à obtenir au minimum 16 images par scène. De plus, les images ont été prises sous différentes conditions météo et différents style photographiques (cf. la figure 2).

L'application initiale de cette collection a été le système Snap2Tell (Lim *et al.,* 2007), dédié à la recherche d'information touristique utilisant des appareil mobiles (par exemple un assistant personnel électronique ou un téléphone portable). Pour les besoins expérimentaux, la collection STOIC-101 est divisée en deux sous-ensembles : l'ensemble d'apprentissage qui contient 3189 images (82.8% de la collection) et l'ensemble de test composé de 660 images (17.15% de la collection). En moyenne, le nombre d'images par classe est de 31,7 pour l'apprentissage et de 6,53 pour le test. Dans l'ensemble de test, le nombre minimum d'image est de 1 et le maximum de 21. Le ratio entre le nombre d'images pour l'apprentissage et le test varie de 12% à 60%. Comme un utilisateur peut prendre une ou plusieurs images de la même scène afin de poser une requête au système de recherche d'information, nous avons considéré plusieurs scénarios d'utilisation :

Figure 2. *Extraits de la base d'images STOIC-101*

1) entraîner le système sur des images isolées et traiter des requêtes d'images isolées ;

2) entraîner le système sur des images isolées, et traiter des requêtes composées d'un groupe d'images de la même scène concaténées ;

3) entraîner le système sur un groupe d'images de la même scène (en les concaténant), et traiter des requêtes d'images isolées ;

4) entraîner le système sur un groupe d'images de la même scène, et traiter des requêtes composées d'un groupe d'images de la même scène.

Le tableau 1 résume ces différents scénarios (une scène correspond à un groupe, toutes les images d'un groupe étant concaténées pour former un seul élément). Notons que

	Entraînement par IMAGE (I)	Entraînement par SCENE (S)
Requête par IMAGE (I)	√	√
Requête par SCENE (S)	√	√

Tableau 1. *Résumé des experimentations sur la collection STOIC-101*

certaines images de la collection ont été remises dans leur orientation correcte (en portrait ou en paysage).

3.2. *Indexation des images avec des concepts et des relations spatiales*

Plusieurs études sur la collection STOIC ont montré que la couleur joue un rôle prédominant, et qu'elle doit être privilégiée par rapport aux autres caractéristiques comme des caractéristiques de bordures ou de texture (Lim *et al.*, 2007). De plus, les caractéristiques de couleurs ont l'avantage d'être extraites facilement et rapidement. Pour ces raisons, nous avons choisi dans ce travail de nous baser sur des caractéristiques de couleurs RGB et HSV. Nous utilisons cependant une méthode de référence sur d'autres caractéristiques visuelles. Cette méthode, appelée *SIFT-couleur*, est similaire aux approches basées sur les points SIFT, largement utilisées et décrites dans (Lowe, 2004). Tout d'abord, les points d'intérêt sont détectés, et pour chaque point un histogramme HSV avec 32 dimensions par canal est utilisé pour indexer le bloc centré sur ce point. Les images de test sont alors comparées avec les images d'entraînement par une distance euclidienne, l'image la plus proche étant utilisée pour catégoriser l'image de test (cette approche revient donc à utiliser une classification de type plus proche voisin avec distance euclidienne).

Pour valider notre méthodologie, nous avons exploré différentes approches pour diviser chaque image en régions, et assigner à chaque région un concept. Pour la division des images en régions, nous avons retenu :

1) Une division à grain fin dans laquelle une région correspond à un pixel (cette approche donne, en moyenne, 86400 régions par image dans la collection). Nous désignons cette division par *gf*, pour grain fin ;

2) Une division "grain moyen", dans laquelle des blocs de 10x10 pixels sont utilisés, le pixel central étant le représentant de la région (cette division donne en moyenne 864 régions par image). Nous appelons cette approche *gm*, pour grain moyen ;

3) Une division grossière, dans laquelle une image est divisée en 5x5 blocs de taille égale. Cette division est appelée *gg*, pour grain grossier.

Pour les divisions *gf* et *gm*, nous avons respectivement quantifié chaque canal RGB et HSV en 8 classes de taille égale (de 0 à 64, etc.). Cela conduit à un vecteur binaire de 512 (8x8x8) dimensions pour une région. Chaque dimension correspond à un concept (défini en fonction des classes des histogrammes), pour lequel chaque dimension correspond à la présence (1) ou l'absence (0) du concept dans la région. L'image globale

est alors indexée par la somme des vecteurs de toutes ses régions. Nous désignerons ces approches par *gf-ConPred* pour “division *gf* avec des concepts préfédinis”, et *gm-ConPred* pour “division *mg* avec des concepts préfédinis”. Ces approches nous servirons de référence pour valider la méthode de regroupement proposée en section 2 pour identifier les concepts de la collection. Dans ce ce que nous venons de décrire, les concepts sont définis arbitrairement au travers des classes des histogrammes, alors qu’en section 2 ils sont définis par regroupement non supervisé.

Pour les divisions *gm* (de nouveau) et *gg*, nous regroupons les vecteurs de caractéristiques HSV de toutes les régions en $K = 500$ classes avec l’algorithme des *K-moyennes*. Cela fournit pour chaque région une affectation stricte de chaque région à un concept. L’ensemble des concepts pondérés W_C est alors obtenu en comptant combien de fois un concept apparaît dans une image. Le choix $K = 500$ est motivé par le fait que nous voulons une certaine granularité pour le nombre de concepts représentant les images. Avec trop peu de concepts, on court le risque de ne pas représenter des différences importances entre les images, alors qu’un nombre trop grand de concepts risque de rendre différentes des images qui sont similaires. Nous appelons les indexations obtenues de cette façon *gm-ConAuto* et *gg-ConAuto*, respectivement pour “division *gm* avec concepts automatiquement générés” et “division *gg* avec concepts automatiquement générés”.

De plus, pour les méthodes *gm-ConAuto* et *gg-ConAuto*, nous avons extrait les relations spatiales entre concepts décrites précédemment (*a_gauche* et *au_dessus*), et nous comptons le nombre d’occurrences de ces relations entre deux concepts donnés afin de les pondérer. La dernière étape fournit un graphe entier pour représenter les images. Nous appelons dans la suite ces deux méthodes *gm-ConAuto-Rel* et *gg-ConAuto-Rel*. Ces deux approches suivent donc le principe décrit en figure 1.

Enfin, pour classifier les images requêtes dans l’une des 101 scènes, nous avons utilisé pour toutes les méthodes d’indexation le modèle de langue pour graphe visuel présenté en section 2. Cela revient à utiliser un classifieur 1-PPV, avec la “similarité” définie par l’équation 1 et ses développements. Quand il n’y a pas de relation, le terme $P(w_E^q|G_d)$ vaut 1 (cf. équation 2), il en résulte que seuls les concepts sont utilisés pour comparer les images.

3.3. *Résultats Expérimentaux*

Les performances des différentes méthodes proposées ont été évaluées d’après le taux de reconnaissance, par image ou par scène. Ce taux est défini comme le ratio d’images (ou de scènes) correctement classifiées :

$$\text{RecoImage} = \frac{TP_i}{N_i}\,,\ \text{RecoScene} = \frac{TP_s}{N_s}$$

où TP_i, resp. TP_s, représente le nombre d’images (resp. scènes) classifié correctement. N_i est le nombre total d’images de test (i.e. 660 images), et N_s est le nombre total de scènes (i.e. 101).

Entraînement	Requête	*gf-ConPred*	*gm-ConPred*	*gm-ConAuto-Rel*	*gg-ConAuto-Rel*
I	I	0.687	0.670	**0.809**	0.551
I	S	0.653	0.650	**0.851**	0.762
S	I	0.409	0.402	0.594	**0.603**
S	S	0.940	0.940	**1.00**	0.920

Tableau 2. *Comparaison globale des differentes méthodes (meilleurs résultats en gras)*

Le tableau 2 présente les résultats que nous avons obtenus en utilisant des concepts prédéfinis et les concepts identifiés automatiquement. Nous constatons que les concepts groupés automatiquement avec un grain moyen fournissent de meilleurs résultats (la différence avec la division à grain grossier pour le scénario S-I étant marginale). Pour le scénario I-I, la méthode *couleur-SIFT* décrite précédemment atteint seulement un taux de 0,425. Ceci montre que pour cette collection notre choix de se focaliser uniquement sur des caractéristiques de couleur semble adéquate[2]. Un autre élément intéressant à souligner est que la division à grain grossier n'aide pas à généraliser par rapport à l'approche à grain moyen. En particulier, les scénarios S-I et I-S correspondent en fait à une utilisation dégénérée du système car les ensembles d'entraînement et de test sont de nature différente. Dans ces cas, il est préférable de s'abstraire d'une description très fidèle des images, afin de généraliser correctement à de nouvelles données de test. L'évolution du taux de reconnaissance des méthodes *gf-ConPred* and *gg-ConAuto-Rel* illustre ce point : les taux de reconnaissance pour les scénarios I-S et S-I sont meilleurs que ceux du scénario I-I pour *gg-ConAuto-Rel*, alors qu'il est plus mauvais pour *fg-ConPred* (ce dernier point se vérifiant également pour la méthode *mg-ConPred*, même si la différence est moins marquée, comme l'on s'y attendait). La méthore *fg-ConPred*, fondée sur une indexation qui est très fidèle à l'image originale, n'est capable de bien généraliser pour aucun des usages.

Ceci étant dit, il y a une différence importante entre les scénarios I-S et S-I : le système traite des requêtes avec davantage d'information dans le scénario I-S que dans S-I. Cette différence a un impact sur les performances pour chaque méthode : les résultats sont moins bons pour le scénario S-I que pour tous les autres, et cela pour toutes les méthodes utilisées. Nous pensons que c'est là l'explication du fait que les résultats obtenus pour la méthode *gm-ConAuto-Rel* pour S-I sont moins bons que pour I-I. Il semble qu'il y ait un plateau pour le scénario S-I autour de 0,6. Nous comptons à l'avenir explorer ce phénomène plus avant.

Nous avons aussi évalué l'utilité des relations spatiales, en comparant les résultats entre les méthodes avec et sans ces relations. Les résultats sont présentés dans le ta-

2. Sur STOIC-101, en utilisant le même découpage entrainement/test, D.-D. Le et S. Satoh, du National Institute of Informatics au Japon, ont obtenu un taux de reconnaissance de 0,744 en utilisant un système à vecteurs de support avec des caractéristiques fondées sur des moments de couleur, des motifs binaires locaux et d'orientation de bordures (communication personnelle).

Entraînement	Requête	*gm-ConAuto*	*gm-ConAuto-Rel*	*gg-ConAuto*	*gg-ConAuto-Rel*
I	I	0.789	**0.809** (+2.5%)	0.484	0.551 (+13.8%)
I	S	0.822	**0.851** (+3.6%)	0.465	0.762 (+63.8%)
S	I	0.529	0.594 (+12.3%)	0.478	**0.603** (+26.1%)
S	S	**1.00**	**1.00**	0.891	0.920 (+3.2%)

Tableau 3. *Impact des relations spatiales sur les performances (meilleurs resultats en gras ; amélioration relative par rapport à la méthode sans relation entre parenthèse)*

bleau 3. Comme nous le constatons, l'utilisation de relations spatiales améliore dans tous les cas les résultats, sauf dans le scénario S-S avec la division *gm*. Ce résultat justifie l'approche modèle de langue sur graphes, avec détection automatique de concepts et prise en compte de relations spatiales, développée dans la section 2.

4. Conclusion

Nous avons introduit dans cet article une nouvelle méthodologie pour indexer des images par des graphes de concepts reliés entre eux. Nous avons de plus proposé un modèle fondé sur le modèle de langue utilisé en recherche d'information pour apparier de tels graphes. Les graphes que nous utilisons capturent les relations spatiales entre les concepts associés à des régions dans des images. Du point de vue formel, notre modèle s'inscrit dans les approches à base de modèle de langue, et étend un certain nombre de travaux antérieurs. A un niveau pratique, l'utilisation de régions et de concepts associés permet un gain en généralité lors de la description des images, une généralité qui est bénéfique lorsque l'usage du système diffère de son environnement d'entraînement. Ceci a de grandes chances de se produire en pratique dès lors que l'on considère des collections où une ou plusieurs images peuvent être utilisées pour représenter une scène.En fonction de l'entraînement réalisé (fondé sur une image ou un ensemble d'images pour une catégorie), les résultats du système varieront.

Les expérimentations menées ont visé à estimer la validité de notre approche par rapport à ces éléments. Nous avons en particulier montré que l'utilisation de relations spatiales conduit à une amélioration significative des résultats. Le modèle proposé est capable de rechercher avec qualité des images et des ensembles d'images représentés par des graphes. De plus, nous avons montré la qualité de notre procédure pour extraire automatiquement les concepts des images, par l'utilisation d'une approche de partionnement classique (K-moyennes). Ces résultats, qui sont les meilleurs présentés sur cette collection, suggèrent qu'une division à grain moyen des images, combiné avec l'utilisation de relations spatiales, constitue une bonne stratégie pour décrire et rechercher des images.

Dans le futur, nous allons utiliser le modèle de graphe décrit ici avec différentes mesures de divergences. Le cadre que nous avons étudié ici est relié à la divergence de Kullback-Leibler. Cependant, la divergence de Jeffrey, utilisée avec succès sur des

collections d'images, pourrait avantageusement remplacer celle de Kullback-Leibler. Nous voulons également étudier les différents couplages entre grain fin, moyen et grossier, avec l'idée d'obtenir une unique représentation utilisable dans tous les cas.

Acknowledgement

Ce travail a été partiellement financé par l'Agence Nationale de la Recherche (ANR-06-MDCA-002).

5. Bibliographie

Barnard K., Duygulu P., Forsyth D., de Freitas D., Blei D., Jordan M. J., « Matching Words and Pictures », *Journal of Machine Learning Research*, vol. 2003, n° 3, p. 1107-1135, 2003.

Chang Y., Ann H., Yeh W., « A unique-ID-based matrix strategy for efficient iconic indexing of symbolic pictures », *Pattern Recognition*, vol. 33, n° 8, p. 1263-1276, 2000.

Fergus R., Perona P., Zisserman A., « A sparse object category model for efficient learning and exhaustive recognition », *Conference on Computer Vision and Pattern Recognition*, 2005.

Gao J., Nie J.-Y., Wu G., Cao G., « Dependence language model for information retrieval », *In SIGIR '04 : Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 170-177, 2004.

Gosselin P., Cord M., Philipp-Foliguet S., « Kernels on bags of fuzzy regions for fast object retrieval », *International conference on Image Processing*, 2007.

Iyengar G., Duygulu P., Feng S., Ircing P., Khudanpur S. P., Klakow D., Krouse M. R., Manmatha R., Nock H. J., Petkova D., Pytlik B., Virga P., « Joint Visual-Text Modeling for automatic Retrieval of Multimedia Documents », *In ACM Multimedia, Singapore*, p. 21-30, 2005.

Kennedy L., Naaman M., Ahern S., Nair R., Rattenbury T., « How flickr helps us make sense of the world : context and content in community-contributed media collections », *In Proceedings of the 15th international Conference on Multimedia*, p. 631-640, 2007.

Lim J., Li Y., You Y., Chevallet J., « Scene Recognition with Camera Phones for Tourist Information Access », *In ICME 2007, International Conference on Multimedia & Expo*, 2007.

Lowe D. G., « Distinctive image features from scale-invariant keypoints », *Journal of Computer Vision*, vol. 60, n° 2, p. 91-110, 2004.

Maisonnasse L., Gaussier E., Chevallet J., « Revisiting the Dependence Language Model for Information Retrieval », *poster SIGIR 2007*, 2007.

Maisonnasse L., Gaussier E., Chevallet J., « Multiplying Concept Sources for Graph Modeling », *In C. Peters, V. Jijkoun, T. Mandl, H. Muller, D.W. Oard, A. Peñas, V. Petras, D. Santos, (Eds.) : Advances in Multilingual and Multimodal Information Retrieval. LNCS #5152. Springer-Verlag.*, 2008.

Mulhem P., Debanne E., « A framework for Mixed Symbolic-based and Feature-based Query by Example Image Retrieval », *International Journal for Information Technology*, vol. 12, n° 1, p. 74-98, 2006.

Ounis I., Pasca M., « RELIEF : Combining Expressiveness and Rapidity into a Single System », *In SIGIR '98 : Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, p. 266-274, 1998.

Papadopoulos T., Mezaris V., Kompatsiaris I., Strintzis M. G., « Combining Global and Local Information for Knowledge-Assisted Image Analysis and Classification », *EURASIP Journal on Advances in Signal Processing*, 2007.

Platt J. C., Czerwinski M., Field B. A., « PhotoTOC : Automatic Clustering for Browsing Personal Photographs », *Proc. Fourth IEEE Pacific Rim Conference on Multimedia*, 2003.

Ponte J. M., Croft W. B., « A language modeling approach to information retrieval », *In SIGIR '98 : Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, p. 275-281, 1998.

Smith J. R., Chang S. F., « VisualSEEk : a fully automated content-based image query system », *In Proceedings of the Fourth ACM international Conference on Multimedia*, p. 87-98, 1996.

Song F., Croft W. B., « general language model for information retrieval », *CIKM'99*, p. 316-321, 1999.

Srikanth M., Srikanth R., « Biterm language models for document retrieval », *Research and Development in Information Retrieval*, p. 425-426, 2002.

Yuan J., Li J., Zhang B., « Exploiting spatial context constraints for automatic image region annotation », *In Proceedings of the 15th international Conference on Multimedia*, p. 25-29, 2007.

Zhai C., Lafferty J., « A study of smoothing methods for language models applied to information retrieval », *ACM Trans. Inf. Syst.*, vol. 22, n° 2, p. 179-214, 2004.